ORIGINAL ARTICLE

Validity, reliability and stability of the portable Cortex Metamax 3B gas analysis system

D. J. Macfarlane · P. Wong

Abstract This study investigated the performance of the portable Cortex Metamax 3B (MM3B) automated gas analysis system during both simulated and human exercise using adolescents. Repeated measures using a Gas Exchange System Validator (GESV) across a range of simulated metabolic rates, showed the MM3B to be adequately reliable (both percentage errors, and percentage technical error of measurements <2%) for measuring expired ventilation (V_E), oxygen consumption (VO_2), and carbon dioxide production (VCO_2). Over a 3 h period, the MM3B was shown to be acceptably stable in measuring gas fractions, as well as V_E, VO_2, and VCO_2 generated by the GESV, especially at moderate and high metabolic rates (drifts <2% and of minor physiological significance). Using eight healthy adolescents during rest, moderate, and vigorous cycle ergometry, the validity of the MM3B was tested against the primary criterion Douglas bag method (DBM) and a secondary criterion machine known to be accurate, the Jaeger Oxycon Pro system. No significant errors in V_E were noted, yet the MM3B significantly overestimated both VO_2 and VCO_2 by approximately 10–17% at moderate and vigorous exercise as compared to the DBM and at all exercise levels compared to the Oxycon Pro. No significant differences were seen in any metabolic variable between the two criterion systems (DBM and Oxycon Pro). It is concluded the MM3B produces acceptably stable and reliable results, but is not adequately valid during moderate and vigorous exercise without some further correction of VO_2 and VCO_2.

Communicated by Susan A. Ward.

D. J. Macfarlane (✉) · P. Wong
Institute of Human Performance, The University of Hong Kong,
Pokfulam, Hong Kong
e-mail: djmac@hku.hk

Introduction

The measurement of oxygen uptake (VO_2) and its associated variables, carbon dioxide production (VCO_2) and expired ventilation (V_E), are now commonly performed in many laboratories around the world in order to assess cardiorespiratory fitness and the metabolic demands of various activities. Traditionally these measurements were undertaken in controlled laboratory conditions via open-circuit calorimetry using developments of the original Douglas bag method (DBM) (Douglas 1911). The advent of automated computerized metabolic gas analysis systems has generally overtaken the time-consuming and skill-dependent DBM in most labs (Macfarlane 2001). The DBM can be used in field trials (Daniels 1971), but it remains very limited due to its bulk, added air-resistance, and inability to easily undertake sequential measurements (Durnin and Passmore 1967). Consequently, a number of portable systems have been designed to acquire metabolic gas measurements during field studies.

One of the earliest portable systems was the fully mechanical Max-Planck respirometer, developed during the Second World War and often referred to as the Kofranyi–Michaelis respirometer (Johnson et al. 1967) after the authors of an early publication (Kofranyi and Michaelis 1949). The early automated portable systems (e.g., "Oxycon", "Oxylog" and "K2") were restricted to only the measurement of V_E and VO_2 (not VCO_2) (Macfarlane 2001). Modern technologies now permit portable systems weighing less than 2 kg to possess virtually all the data collection powers of their lab-based counterparts (often recording or telemetering breath-by-breath and heart rate

data). One such system is the Metamax 3B (MM3B) system (Cortex, Leipzig, Germany), that is also marketed as the VmaxST in many countries, which supersedes the earlier Metamax I and II models that were shown to be valid and reliable (Medbo et al. 2002).

To be able to be used in discipline-specific field studies, these portable gas analysis systems have to be not only small and unobtrusive, but also reliable and valid. There is a lack of consensus on not only the level of precision and accuracy expected for measures of VO_2 and V_E (Macfarlane 2001), but also which method is most appropriate to assess both reliability and validity. For example, studies have reported the reliability of the MM3B by comparing data acquired from human participants twice on different days (Perkins et al. 2004), yet this analysis will include not only the technical error but also the biological within-subject variation. Only reliability comparisons of dual measurements using a gas exchange simulator will remove the within-subject variation, and this has been reported recently with the MM3B (Vogler et al. 2010). Three validation studies on the MM3B have been performed against the DBM (Brehm et al. 2004; Prieur et al. 2003; Vogler et al. 2010), and despite its limitations the DBM is often considered to be the criterion (Hodges et al. 2005), especially when precise metabolic calibrators are not available (Gore et al. 1997). Yet other validation studies on the MM3B have only used a previously validated automated system as its criterion measure (Laurent et al. 2008; Perkins et al. 2004), a method that has been questioned as not being a true gold standard (Meyer et al. 2005). Ideally, any metabolic gas analysis system, such as the MM3B, should have its reliability reported using a gas exchange simulator and its validity reported compared to at least the DBM. Only one extensive study has published such data on the MM3B (Vogler et al. 2010); however, this study focused on elite athletes at relatively high levels of performance and did not include resting or light activity. In order to determine if the MM3B could be applied to the common non-elite populations such as children, validity and reliability trials were needed at much lower physiological levels.

Furthermore, it is important for any metabolic gas analysis system to also demonstrate that, once calibrated, it does not drift significantly (i.e., is stable) over the usual data-acquisition period. For measurements taken in or near a laboratory this may only require determining the stability over a 20 min period (Prieur et al. 2003), but for some studies, the calibration of a portable system in the field may be problematic due to the need for carrying bulky calibration equipment, including large pressurized bottles of alpha-standard calibration gases. Thus, some portable systems may need to be initially calibrated in a laboratory and then transported and used at a significant time later for field studies. The identification of how each gas analyzer and resultant VO_2 and VCO_2 drifts in these situations is

rarely reported (Atkinson et al. 2005). Only one study has reported on the stability of the Cortex Metamax 3B/VmaxST system, but this was for only a maximum of 20 min (Prieur et al. 2003) and many field studies may exceed this duration. Furthermore, it is important to match the type of validation as to how the testing system will be used (Unnithan et al. 1994), and although many automated metabolic systems are used on pediatric/adolescent populations, few systems have been actually validated using these groups (Unnithan et al. 1994).

The aim of this study was therefore to report on the suitability of the MM3B to measure variables over non-elite physiological ranges, especially those commonly found in pediatric/adolescent or elderly groups, as well as its ability to be used in prolonged field situations. We therefore studied the performance of the MM3B system for (1) reliability, using a commercially available gas exchange simulator; (2) stability/drift over a 3 h period; (3) validity, compared to both the criterion DBM as well as a previously validated gas analysis system (Jaeger Oxycon Pro) using an adolescent sample.

Methods

Participants

Eight healthy young volunteers (3 boys, 5 girls) were recruited with the following characteristics (mean $\pm$ SD): age 12.9 ± 3.6 years; height 150.1 ± 13.1 cm; mass 44.6 ± 11.5 kg. All subjects, as well as their legal guardians, provided a written informed consent after the project was approved by the Research Ethics Committee of the University of Hong Kong.

Equipment

Metamax 3B (breath-by-breath system)

The MM3B is a portable metabolic system composed of a measurement module and a battery module. These two parts are of the same size ($120 \times 110 \times 45$ mm) and designed to be worn on the chest with a harness, with a total weight of 1.40 kg. The MM3B measures volume using a bidirectional digital turbine. A 60 cm length of Nafion/Permapure sampling tube is attached to the turbine to permit analysis of the O_2 and CO_2 concentrations using an electrochemical cell and an infrared analyzer, respectively. VO_2 and VCO_2 were calculated using standard metabolic algorithms (Wasserman et al. 1999) employing the Haldane transformation, but with F_IO_2 and F_ICO_2 continuously measured, rather than assumed to be constant, in order to correct for changes in ambient conditions.

The breath-by-breath data of respiratory volume and gas concentrations can be stored in on-board memory for later downloading to a PC, or sent immediately via telemetry to a PC. The system was tested using Metasoft 3 software, version 3.7.0 SR2.

Prior to using, the system was turned on for at least 20 min, and then calibrated prior to every test according to the manufacturer's recommendations. This involves first calibrating the gas analysers by using a reference gas (14.97% O_2, 4.96% CO_2, balance N_2: $\pm 0.02\%$ absolute, Hong Kong Specialty Gases), and then verifying the calibration against ambient air. Secondly, a volume calibration was performed using a standardized 3-L syringe (5530 series, Hans Rudolph, Inc., MO, USA). For avoiding potential gas leakages known to be problematic with facemasks, all participants wore a nose clip and had a mouthpiece attached to the MM3B turbine.

Gas Exchange System Validator (GESV)

The GESV (MedGraphics, MA, USA; a similar, but updated, GESV is now sold by Vacumed, CA, USA) was a mechanical gas exchange system simulating human breathing and has a reported accuracy in producing VO_2 and VCO_2 of $\pm 2\%$ (Huszczuk et al. 1990). When a gas of known CO_2 concentration (recommended 21.00%) was added to the inspirate, the GESV expired gas of constant expired fractions at ambient temperature and pressure that could be used to simulate a range of VO_2 and VCO_2. The GESV could be adjusted so that it simulated breathing over a wide range of tidal volumes (0.5, 1.0, 1.5, 2.0, 2.5 and 3.0 L) at various respiratory rates (low = 10 breaths min^{-1}; medium = 20 breaths min^{-1}; and high = 40 breaths min^{-1}). These resulted in minute ventilations ranging from 5 to 104 L min^{-1}; VO_2 ranging from 0.30 to 2.81 L min^{-1}; and VCO_2 ranging from 0.29 to 2.69 L min^{-1}.

Douglas bag method (DBM)

With each participant wearing a nose clip and mouthpiece attached to a Radiax valve (dead space ~ 28 ml), all expired gases were collected in 250 L Douglas bags (WE Collins, Braintree, USA) using a 1 m length of Collins spiral tubing (38 mm ID) and a Collins 3-way stopcock. The mixed expirates were analyzed within 15 min using an S-3A oxygen and CD-3A carbon dioxide analyzer (Applied Electrochemistry, Sunnyvale, CA: now AEI Technologies) that had been previously calibrated using two alpha/primary-reference gases (26.13% O_2 and 0.00% CO_2; 13.94% O_2 and 5.96% CO_2; all gases $\pm 0.02\%$ absolute, Hong Kong Specialty Gases) and checked against ambient air (analyser linearity was checked using 0.00, 13.94 and 26.13% O_2,

and 0.00, 4.96 and 5.96% CO_2 as well as ambient air). The volume of the Douglas bag was measured by a dry gas meter (Harvard, USA), with the aid of a vacuum pump on the exit port. The temperature of the expired gas was monitored with a telethermometer (YSI, Ohio, USA) placed at the outlet of the gas meter for later correction to "standard temperature pressure dry" (STPD), and "body temperature pressure saturated" (BTPS). All bags were checked for an absence of leaks and diffusion (no change in volume or composition of mixed expirate noted over a 30 min period), and each one flushed with expired gases before use to reduce the dilution effect of dead-space gas trapped in the bag and any rigid tubing.

Jaeger Oxycon Pro

The Oxycon Pro (Jaeger—now CareFusion, Germany; running JLab Software version 4.66.0) was used in its "mixing-chamber mode" with participants, wearing a nose clip, breathing via a mouthpiece and Radiax valve. The system was turned on for at least 30 min prior to use, and then fully calibrated for gases (14.00% O_2, 6.00% CO_2: $\pm 0.02\%$ absolute, Hong Kong Specialty Gases) and volumes before every test according to the manufacturer's recommendations.

Studies

Study 1—reliability

Reliability trials of V_E, VO_2 and VCO_2 measured by the MM3B were examined by attaching a known CO_2 gas supply (20.62%) to the inspired port of the GESV as recommended by the manufacturer. Each trial consisted of the GESV working at 1.0 L at low respiratory rate, 1.5 L at medium respiratory rate, and 1.5 L at high respiratory rate, with the GESV inspiratory/expiratory port connected directly in-series to the MM3B turbine. Each trial was repeated twice at each level of V_E, VO_2, and VCO_2 in order to assess the reliability of this measure, all during the same day, with re-calibration of the MM3B between each trial.

Study 2—stability

To measure the stability/drift of two components of the MM3B (a) the gas analysers; a known gas (15.83% O_2 and 4.05% CO_2) was introduced to the sampling line of the MM3B at 0, 20, 40, 60, 120 and 180 min; (b) the full system; simulated VO_2 and VCO_2 were introduced by attaching a 20.62% CO_2 gas to the inspired port of the GESV as recommended by the manufacturer. A trial

consisted of the GESV working at 1.0 L at low respiratory rate, 1.5 L at medium respiratory rate, and 1.5 L at high respiratory rate, with the GESV inspiratory/expiratory port connected directly to the MM3B turbine for 30 complete "breaths" at 20-min intervals until 180 min had elapsed. The 180-min time period was considered as the maximum period the MM3B would be used during field measures once it had been calibrated prior to data collection.

Study 3—validity

The eight participants performed one trial comprising stages of resting and incremental cycling exercise (Corival 400, Lode, The Netherlands). Each exercise stage lasted 13 min and measurements were made in-parallel (sequentially) with the DBM, the MM3B, and an Oxycon Pro system. The Oxycon Pro system acted as a "secondary criterion" as it has been shown to be a valid metabolic system in its mixing-chamber mode (Foss and Hallen 2005). The DBM collections of expired gas were used as the primary criterion and the order of measurement system followed a counterbalanced Latin-Square process so as to avoid any order effect (Bradley 1958). The expired gas collected by the DBM was immediately measured by the calibrated S-3A oxygen and CD-3A carbon dioxide analyzers and then passed through the dry gas meter whose accuracy (0.8% error) had been previously checked using multiple pumps of a 3-L calibration syringe (Hans Rudolph).

Each trial using the DBM, MM3B, and the Oxycon Pro began with a 10-min rest period, with gas collection over the final 5 min. Each participant then pedaled at a constant rate of 50 rpm in three or four stages, starting from 50W and with a 25W or 50W increment according to their perceived fitness and body size. Each exercise stage lasted for a total of 13 continuous minutes. Each participant exercised for 5 min whilst breathing into the first instrument, and then kept pedaling at the same rate for 1 min during which the experimenter quickly changed to the second measurement system. As the participant was already in steady-state exercise, the participant continued exercise for 3 min with the second measurement system, and then the procedure was repeated for the third measurement system. The three for four intensity levels were later classified as being representative of resting, moderate or vigorous exercise. The data used in calculation of all variables was the mean of the final 2 min of measurement with each device, which was considered to be during the steady state (supported by visual inspection of the heart rate). All trials were performed in a quiet laboratory in environmentally stable conditions; any variations in F_IO_2 and F_ICO_2 during DBM collections were noted from the Applied Electrochemistry analysers and appropriate corrections made in the DBM calculations.

Data analysis

Tests to examine any differences between the dependent variables of V_E, VO_2, VCO_2, F_EO_2, and F_ECO_2 included percentage differences, repeated-measures ANOVA, and the Bland–Altman analysis (Bland and Altman 1986). Univariate Intra-Class Correlation Coefficients (ICC), and t tests (for shift in the mean scores) were also used in the reliability tests between repeated measures. Technical Errors of Measurement (TEM) were generated for both the between-device validity comparisons (inter-TEM), as well as within-device reliability comparisons (intra-TEM), following similar procedures reported by Gore et al. (2000). SPSS (8.0) was used for most analyses, with the TEM = $\sqrt{\sum D^2/2N}$ where D is the difference between the pairs of measurement, and N is the number of measurement pairs.

Results

In the reliability trials of the MM3B in measuring V_E, VO_2, and VCO_2 with a re-calibration in-between (Study 1), significant but small differences were found for VO_2 and VCO_2 at all levels of frequencies. As shown in Table 1, the percentage differences between the two trials were consistency less than 2.5%, whilst the average intra-device TEM across all pumping frequencies was 0.2, 1.4 and 1.1% for V_E, VO_2 and VCO_2, respectively. Ideal ICC values ($r = 1.00$) were also obtained by the MM3B in the repeated measurements of V_E, VO_2, and VCO_2.

The stability/drift trials of the MM3B gas analysers (Study 2) showed that when compared to the baseline (0 min), significant but small differences were found for F_EO_2 and F_ECO_2 at each subsequent measurement time (20, 40, 60, 120 and 180 min). The differences ranged from 0.01 to 0.09% for F_EO_2 and −0.05 to −0.16% for F_ECO_2. Table 2 shows the descriptive statistics for the gas concentration in percentages and comparisons between each time point and the baseline by the MM3B. F_ECO_2 immediately dropped nearly 4% at the first 20-min interval and this under-measurement slowly improved until 120 min where only a 1% difference from the baseline existed. Very small, yet statistically significant, increases were also seen in F_EO_2 as it slowly drifted upwards.

Results of the full system analysis of how VO_2 and VCO_2 drifted over time (Study 2, see Table 2) showed that at low metabolic rate, VO_2 started to drift upwards from the baseline at the 60th minute onwards and resulted in an average of 2% difference. A significant drift of VCO_2 was found at the 20th minute and continued to deviate significantly from the baseline at the 60th minute onwards and this resulted in an average of 3% difference. At medium

Table 1 Reliability results, showing the mean $\pm$ SD values, percentage differences, and intra-device TEM (%) of V_E, VO_2 and VCO_2 (L min^{-1}) of repeated trials by the MetaMax 3B across a range of low, moderate, high metabolic rates generated by the Gas Exchange System Validator

	Metabolic rate	Trial 1	Trial 2	% difference[#]	Intra-device TEM (%)
V_E (BTPS, L min^{-1})	Low	9.94 ± 0.06	9.93 ± 0.08	-0.1	0.1
	Medium	29.53 ± 0.28	29.69 ± 0.23	0.5	0.4
	High	58.06 ± 0.61	58.10 ± 0.65	0.1	0.0
VO_2 (STPD, L min^{-1})	Low	0.30 ± 0.01	0.30 ± 0.00	1.9*	1.3
	Medium	1.62 ± 0.01	1.65 ± 0.01	1.8*	1.3
	High	2.74 ± 0.03	2.80 ± 0.03	2.3**	1.6
VCO_2 (STPD, L min^{-1})	Low	0.29 ± 0.01	0.29 ± 0.00	2.5*	1.7
	Medium	1.57 ± 0.02	1.59 ± 0.02	1.2*	0.8
	High	2.66 ± 0.04	2.69 ± 0.04	1.1*	0.8

TEM% technical error of measurement expressed as a percentage of mean

[#] Percentage differences were calculated from mean data using four decimal places, but the data are reported here to two decimal places only

* Significant difference ($p < 0.05$)

** Significant difference ($p < 0.01$) from t test

and high metabolic rates, some values slightly and significantly higher than the baseline VO_2 were found around the 100–120 min mark. For VCO_2 at medium and high metabolic rates, there were few significant or consistent trends to the drift.

The descriptive statistics for differences between the MM3B as compared to the primary criterion (DBM) and the secondary criterion (Oxycon Pro) are shown in Table 3 (Study 3), whilst detailed comparisons for the resting, moderate, and vigorous activity between the DBM and MM3B using data from the Bland–Altman analysis, percentage differences, as well as TEM are shown in Table 4. The combined data for all activities are shown in Fig. 1, where no systematic error with some proportional random error is seen in V_E, whilst both VO_2 and VCO_2 show evidence of both proportional systematic and random errors (Atkinson et al. 2005). No statistically significant differences in V_E were seen between the DBM and MM3B despite the overall TEM% averaging around 11%, and the MM3B underestimating resting V_E by nearly 9%. However, significant differences were seen in both VO_2 and VCO_2 for moderate and vigorous activity, with the MM3B overestimating these variables by more than 10% and producing TEM% values that varied from 9 to 18%. When compared to the Oxycon Pro, the MM3B varied significantly at all levels of VO_2 and VCO_2, but only for moderate V_E (Table 3). Interestingly, Table 3 also shows that when compared to the primary criterion DBM, the secondary criterion Oxycon Pro showed no significant differences in V_E, VO_2 or VCO_2 at rest, moderate or vigorous activity.

Discussion

This is the first study to comprehensively examine the reliability and prolonged stability of the Cortex Metamax 3B over multiple ranges of simulated exercise conditions, as well as performing validity comparisons against both a primary criterion DBM, and a secondary criterion (Oxycon Pro) over metabolic ranges suited to non-elite participants (e.g., adolescents/elderly).

Meyer et al. (2005) reported that despite the need for establishing the reliability of portable gas analysis devices, few studies have addressed this issue; however, three studies appear to have reported data on the reliability of the Metamax 3B. The study by Perkins et al. (2004) examined reliability of the MM3B but used repeated measurements on human participants, which inflates the variability as it combines the relatively large biological error of the participants and the smaller technical error of the machine [their contributions to the total variability has been estimated to be 90% and 10% respectively (Macfarlane 2001)]. Even so, Perkins et al. (2004) reported the MM3B had extremely high single and multiple trial reliabilities and with narrow confidence intervals. Prieur et al. (2003) reported data from a stability trial using a comprehensive gas exchange simulator that showed the MM3B to be very reliable, although data were provided for only a single metabolic rate ($VO_2 = 2.6$ L min^{-1}). Data from the Vogler study (Vogler et al. 2010) also showed that the MM3B to be extremely reliable, with typical errors that ranged from 2.0% (VO_2) and 3.6% (V_E), which were superior to the reliability of their criterion Douglas bag system.

Table 2 Stability/drift results, showing the mean ± SD values measured by MetaMax 3B of F_EO_2, and F_ECO_2, as recorded from the calibration gas bottle, as well as VO_2 and VCO_2 at different metabolic rates (low, medium, high) as simulated by the Gas Exchange System Validator, over a 180-min period

Time (min)	F_EO_2 (%)	F_ECO_2 (%)	VO_2 (L min^{-1})			VCO_2 (L min^{-1})		
			Low	Med	High	Low	Med	High
0	15.84 ± 0.00	4.10 ± 0.02	0.30 ± 0.00	1.67 ± 0.02	2.82 ± 0.03	0.28 ± 0.00	1.57 ± 0.02	2.66 ± 0.04
20	15.85 ± 0.00 (0.1***)	3.94 ± 0.04 (−3.9***)	0.30 ± 0.00 (−0.1)	1.66 ± 0.01 (−0.4)	2.81 ± 0.04 (−0.5)	0.29 ± 0.00 (1.5***)	1.58 ± 0.02 (0.7)	2.68 ± 0.04 (1.1*)
40	15.87 ± 0.00 (0.2***)	3.99 ± 0.03 (−2.7***)	0.30 ± 0.00 (−0.7)	1.67 ± 0.01 (0.1)	2.82 ± 0.03 (−0.2)	0.28 ± 0.01 (0.7)	1.58 ± 0.01 (1.1*)	2.68 ± 0.03 (0.8)
60	15.88 ± 0.00 (0.3***)	3.97 ± 0.03 (−3.2***)	0.31 ± 0.00 (2.0***)	1.67 ± 0.02 (−0.2)	2.83 ± 0.03 (0.3)	0.29 ± 0.00 (3.4***)	1.58 ± 0.02 (0.8)	2.67 ± 0.03 (0.4)
80	–	–	0.31 ± 0.00 (2.8***)	1.68 ± 0.02 (0.6)	2.84 ± 0.03 (0.9)	0.29 ± 0.00 (3.8***)	1.59 ± 0.02 (1.3*)	2.68 ± 0.04 (1.1)
100	–	–	0.31 ± 0.00 (2.7***)	1.68 ± 0.01 (0.6)	2.85 ± 0.03 (1.0*)	0.29 ± 0.00 (3.5***)	1.58 ± 0.01 (0.6)	2.68 ± 0.04 (1.0)
120	15.91 ± 0.00 (0.4***)	4.05 ± 0.03 (−1.2***)	0.31 ± 0.00 (2.6***)	1.69 ± 0.02 (1.1*)	2.84 ± 0.03 (0.8)	0.29 ± 0.00 (3.2***)	1.58 ± 0.02 (1.0)	2.67 ± 0.05 (0.4***)
140	–	–	0.31 ± 0.00 (2.7***)	1.69 ± 0.01 (1.0**)	2.86 ± 0.03 (1.5***)	0.29 ± 0.00 (3.2***)	1.58 ± 0.02 (0.8)	2.69 ± 0.03 (1.3)
160	–	–	0.31 ± 0.00 (2.6***)	1.69 ± 0.02 (1.3**)	2.85 ± 0.03 (1.0*)	0.29 ± 0.00 (2.8***)	1.58 ± 0.02 (0.8)	2.68 ± 0.04 (0.8*)
180	15.93 ± 0.00 (0.6***)	4.05 ± 0.03 (−1.2***)	0.31 ± 0.00 (2.1***)	1.69 ± 0.01 (1.3***)	2.87 ± 0.03 (1.8***)	0.29 ± 0.00 (2.1***)	1.58 ± 0.01 (0.7)	2.69 ± 0.04 (1.1)

The resultant relative percentage error in comparison with baseline, at time = 0, is shown in brackets

Significantly different to time at 0 min (ANOVA)

$Error$ (%) difference between F_ECO_2, F_ECO_2, VO_2 or VCO_2 values as [each time (min)−baseline (0 min)]/value of baseline × 100%, and was calculated from mean data expressed as four decimal places

* $p < 0.05$

** $p < 0.01$

*** $p < 0.001$

The reliability data from our study that used the GESV to simulate a wide range of conditions likely to be experienced in normal field trials (e.g., low, medium, and high metabolic rates), showed that the technical variability of the MM3B measurements was adequately low. The relative percentage errors for V_E, VO_2 and VCO_2 all being typically less than 2% between tests, with the TEM% generally less than 1.5%. These reliability results compare favorably with a 1% relative error generated from a complex automated calibration system (Gore et al. 1997), and is below the TEM reliability limit of 3% as recommended by the Australian Sports Commission (Gore 2000) for these variables.

An important aim of this study was to examine the stability (or resistance to drift) of key variables measured by the MM3B (gas fractions, VO_2, VCO_2) over a 3-h period that would reflect the longest likely time the MM3B would be used in the field after calibration in the lab (e.g., 60 min travel time to destination, 30 min preparation on-site, then allowing up to 90 min of episodic data collection). In

countries like Hong Kong, where some field measures are undertaken in remote locations with no parking for private vehicles, it is often not feasible to transport calibration equipment to the site, necessitating prior calibration in the lab and reliance that the equipment is adequately stable over time. Our analysis, although limited to static laboratory conditions, therefore partially addresses the comment by Atkinson et al. (2005) that insufficient data is available on how stable specific gas analysis systems are, as does the paper by Eriksson et al. (2011).

The results of this study agrees with that of Prieur et al. (2003) that the MM3B shows some statistically significant drift, however, the absolute magnitude of this drift is small. The relative errors appear to be large (2–3%) only when the comparative original value is small (i.e., low denominator in the resting conditions), but under typical moderate-to-vigorous exercise conditions, the absolute error is relatively minor and is seen more in VCO_2 and likely due to a greater drift in the F_ECO_2 measurement. As these drifts in VO_2 and

Table 3 Validity results, showing the mean $\pm$ SD values of physiological variables measured by the three metabolic systems during parallel data collection at rest, moderate, and vigorous cycle exercise

* Significantly different to Douglas bag measurements (ANOVA, $p < 0.05$)

^ Significantly different to Oxycon Pro measurements (ANOVA, $p < 0.05$)

	Activity	Douglas bag	MetaMax 3B	Oxycon Pro
V_E (BTPS, L min^{-1})	Rest	8.72 $\pm$ 1.03	7.93 $\pm$ 1.73	7.72 $\pm$ 1.82
	Moderate	29.89 $\pm$ 8.19	30.61 $\pm$ 9.78^	28.73 $\pm$ 8.31
	Vigorous	66.00 $\pm$ 20.52	66.25 $\pm$ 19.16	62.54 $\pm$ 18.26
VO_2 (STPD, L min^{-1})	Rest	0.27 $\pm$ 0.04	0.30 $\pm$ 0.08^	0.25 $\pm$ 0.06
	Moderate	1.12 $\pm$ 0.33	1.24 $\pm$ 0.41*^	1.12 $\pm$ 0.36
	Vigorous	2.32 $\pm$ 0.46	2.59 $\pm$ 0.52*^	2.37 $\pm$ 0.44
VCO_2 (STPD, L min^{-1})	Rest	0.22 $\pm$ 0.04	0.26 $\pm$ 0.07^	0.20 $\pm$ 0.06
	Moderate	1.03 $\pm$ 0.31	1.17 $\pm$ 0.40*^	1.01 $\pm$ 0.33
	Vigorous	2.25 $\pm$ 0.52	2.64 $\pm$ 0.66*^	2.17 $\pm$ 0.43

Table 4 Validity analyses of the Metamax 3B data against the Douglas bag (primary criterion) during parallel data collection at rest, moderate, and vigorous cycle exercise

Variable	Activity level	Bland–Altman bias (L min^{-1}) Mean $\pm$ SD (95% LOA)	% difference Mean $\pm$ SD	Inter-device TEM (%) Mean
V_E (BTPS, L min^{-1})	Rest	-0.79 ± 1.62 (2.45, -4.04)	-8.8 ± 19.3	17.0
	Moderate	0.72 ± 2.76 (6.24, -4.80)	1.5 ± 9.8	7.0
	Vigorous	0.25 ± 10.02 (20.28, -19.79)	2.2 ± 13.3	9.7
VO_2 (STPD, L min^{-1})	Rest	0.03 ± 0.05 (0.14, -0.07)	10.6 ± 19.3	14.0
	Moderate	0.12 ± 0.14** (0.41, -0.16)	9.7 ± 13.2**	10.9
	Vigorous	0.27 ± 0.16** (0.59, -0.05)	11.8 ± 7.6**	9.4
VCO_2 (STPD, L min^{-1})	Rest	0.04 ± 0.05 (0.14, -0.06)	17.3 ± 21.8	18.0
	Moderate	0.14 ± 0.11** (0.36, -0.08)	12.5 ± 9.8**	10.6
	Vigorous	0.39 ± 0.23** (0.85, -0.07)	17.4 ± 8.1**	12.8

Data shown from Bland–Altman analyses (mean $\pm$ SD of bias, and 95% limits of agreement, LOA), percentage differences between means, and the inter-device TEM (%)

Bias and differences computed for data calculated from the Metamax 3B–Douglas bag

TEM% technical error of measurement expressed as percentage of mean value

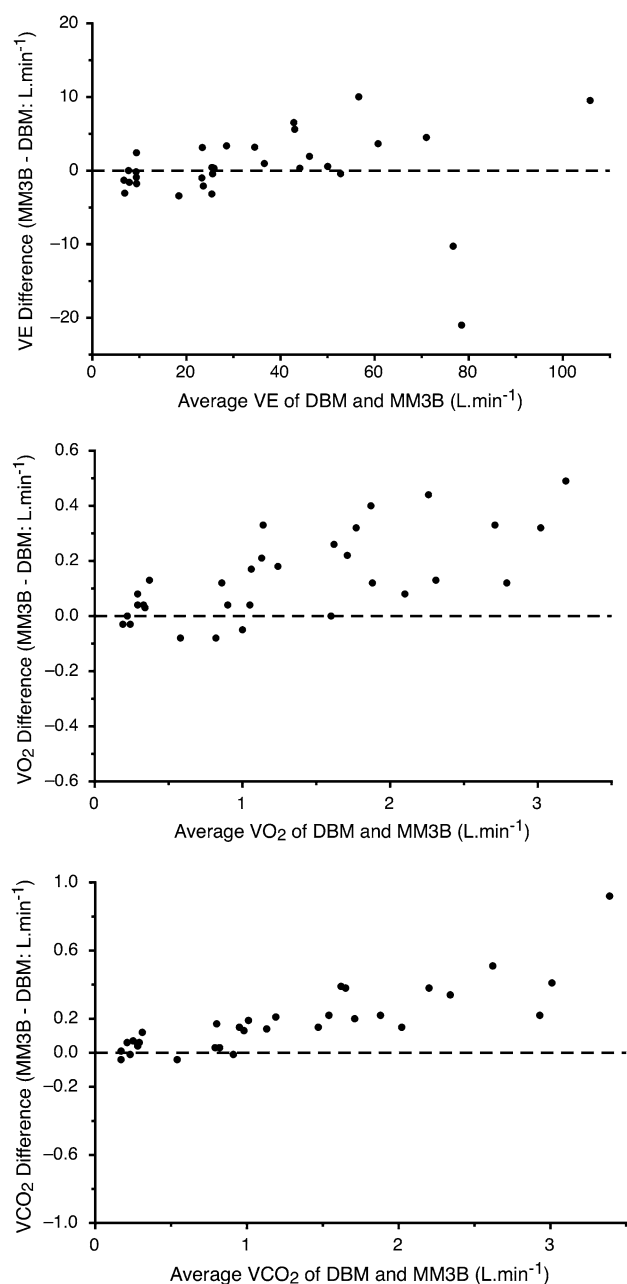** Metamax 3B significantly different to Douglas bag measurements (ANOVA, $p < 0.01$)

Fig. 1 Modified Bland–Altman plots showing agreement between measurements taken across all activities combined (rest, moderate, vigorous) for all eight subjects from the Metamax 3B (MM3B) and criterion Douglas bag method (DBM). *Y* axes showing the absolute differences in V_E (BTPS), VO_2 (STPD) and VCO_2 (STPD) plotted against the average values of both methods (*X* axes)

VCO_2 during simulated intermittent exercise lasting 180 min were all below 2%, there are unlikely to be of physiological importance, to the extent that the MM3B can be considered very stable.

During moderate and vigorous exercise the MM3B significantly overestimated VO_2 and VCO_2, but not V_E, by 10–17% when compared to the primary criterion DBM, but at all VO_2 and VCO_2 values during exercise when

compared to the secondary criterion Oxycon Pro. In comparison, there were no significant differences in VO_2, VCO_2 or V_E across all conditions between the primary (DBM) and secondary (Oxycon) criterion machines, which support a previous study showing the Oxycon Pro's mixing-chamber mode to be very accurate (Foss and Hallen 2005). However, for the key VO_2 variable, previous validation studies on the MM3B have produced inconsistent findings, with both overestimates (Perkins et al. 2004; Vogler et al. 2010) and underestimates (Brehm et al. 2004; Laurent et al. 2008; Prieur et al. 2003) being reported. The relatively large percentage errors reported in our study for VO_2 and V_E by the MM3B are generally higher than those reported in the above validation studies, and also exceed the 4–10% guidelines recommended by some (Brehm et al. 2004; Laurent et al. 2008; Vogler et al. 2010), although these limits are not universally agreed upon (Macfarlane 2001). It is worth noting that some percentage differences (in the means) between the MM3B and DBM in Table 4 were relatively small (e.g., V_E during moderate and vigorous exercise, <2.5%), yet were associated with relatively large TEM% scores of >7%. This was possibly due to large (quasi-symmetric) variation in the pairs of data around the mean, as the TEM% is sensitive to the degree of variability in the data pairs (and systematic error), yet this variability is not reflected in the percentage difference of the mean scores as it is only sensitive to systematic error.

It is unclear why our relative errors in VO_2 and VCO_2 are higher than other MM3B validation studies, as it is unlikely this was due to errors made in the Douglas bag assessments, as there was good agreement between the primary DBM and secondary Oxycon Pro criterions. Nor was greater variation likely to be due to our small number of participants ($n = 8$), as other validation studies have also used comparable numbers ($n = 8$–11) of participants (Brehm et al. 2004; Laurent et al. 2008; Prieur et al. 2003; Vogler et al. 2010). The slow drift upwards in VO_2 and VCO_2 over time by <4% at a low metabolic rate and <2% at the two higher metabolic rates, as reported in Table 2, may partially account for some, but not all, of the error. It would appear that the additional challenges of measurement during dynamic human exercise, such as movement shocks, gas leakages, saliva entrapment, and variations in gas and flow waves as suggested by Prieur et al. (2003), might incur greater measurement errors that are not seen during static testing using a mechanical simulator.

This study contains several limitations. Ideally, validations should be done using a serial method so that all expired gases passes sequentially through the MM3B and then into the Douglas bag; however, as outlined by others (Prieur et al. 2003; Vogler et al. 2010), and confirmed in our lab, this was not possible due to interference during simultaneous measurements, and instead we used separate

(parallel) trials. The stability/drift analysis was performed only under stable laboratory conditions and without the device being subjected to regular movements and environmental changes as might occur during transport outdoors to a remote venue. It is unknown how changing environmental conditions might influence the validity and reliability of the MM3B. As we were interested in evaluating the MM3B for prolonged field studies on children, the participants we recruited for the validation phase of this study were not highly trained athletes, hence the upper range of their metabolic responses even during vigorous exercise are unlikely to be cover the range necessary to study elite athletes.

Summary

In many situations the MM3B will be used in field studies and when used at remote sites the results of this study suggest this device remains acceptably stable (significant variations were of minor physiological importance) for periods up to 3 h. The MM3B is also very reliable, but appears to be insufficiently valid when measuring VO_2 and VCO_2 during moderate and high intensities (evidence of proportional systematic and random errors), although these errors may be mitigated using a simple linear regression equation.

References

Atkinson G, Davison RC, Nevill AM (2005) Performance characteristics of gas analysis systems: what we know and what we need to know. Int J Sports Med 26(Suppl 1):2–10

Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 327:307–310

Bradley JV (1958) Complete counterbalancing of immediate sequential effects in a Latin Square design. J Am Stat Assoc 53:525–528

Brehm MA, Harlaar J, Groepenhof H (2004) Validation of the portable VmaxST system for oxygen-uptake measurement. Gait Posture 20:67–73

Daniels J (1971) Portable respiratory gas collection equipment. J Appl Physiol 31:164–167

Douglas CG (1911) A method for determining the total respiratory exchange in man. J Physiol 42:17–18

Durnin JVGA, Passmore R (1967) Energy, work and leisure. Heinemann Educational, London

Eriksson JS, Rosdahl H, Schantz P (2011) Validity of the Oxycon Mobile metabolic system under field measuring conditions. Europ J Appl Physiol. doi: 10.1007/s00421-00011-01985-00421

Foss O, Hallen J (2005) Validity and stability of a computerized metabolic system with mixing chamber. Int J Sports Med 26:569–575

Gore CJ (2000) Quality assurance in exercise physiology laboratories. In: Gore CJ (ed) Physiological testing for elite athletes (Australian Sports Commission). Human Kinetics, Champaign, pp 3–11

Gore CJ, Catcheside PG, French SN, Bennett JM, Laforgia J (1997) Automated VO_{2max} calibrator for open-circuit indirect calorimetry systems. Med Sci Sports Exerc 29:1095–1103

Hodges LD, Brodie DA, Bromley PD (2005) Validity and reliability of selected commercially available metabolic analyzer systems. Scand J Med Sci Sports 15:271–279

Huszczuk A, Whipp BJ, Wasserman K (1990) A respiratory gas exchange simulator for routine calibration in metabolic studies. Eur Respir J 3:465–468

Johnson RE, Robbins F, Schilke R, Mole P, Harris J, Wakat D (1967) A versatile system for measuring oxygen consumption in man. J Appl Physiol 22:377–379

Kofranyi E, Michaelis HF (1949) Ein tragbarer Apparat zur Bestimmung des Gasstoffwechsels. Arbeitsphysiologie 11:148–150

Laurent CM, Meyers MC, Robinson CA, Strong LR, Chase C, Goodwin B (2008) Validity of the VmaxST portable metabolic measurement system. J Sports Sci 26:709–716

Macfarlane DJ (2001) Automated metabolic gas analysis systems: a review. Sports Med 31:841–861

Medbo JI, Mamen A, Welde B, von Heimburg E, Stokke R (2002) Examination of the Metamax I and II oxygen analysers during exercise studies in the laboratory. Scand J Clin Lab Invest 62:585–598

Meyer T, Davison RC, Kindermann W (2005) Ambulatory gas exchange measurements—current status and future options. Int J Sports Med 26(Suppl 1):19–27

Perkins CD, Pivarnik JM, Green MR (2004) Reliability and validity of the VmaxST portable metabolic analyzer. J Phys Activ Heal 1:413

Prieur F, Castells J, Denis C (2003) A methodology to assess the accuracy of a portable metabolic system (VmaxST). Med Sci Sports Exerc 35:879–885

Unnithan VB, Wilson J, Buchanan D, Timmons JA, Paton JY (1994) Validation of the Sensormedics (S2900Z) metabolic cart for pediatric exercise training. Can J Appl Physiol 19:472–479

Vogler AJ, Rice AJ, Gore CJ (2010) Validity and reliability of the Cortex MetaMax3B portable metabolic system. J Sports Sci 28:733–742

Wasserman K, Hansen J, Sue D, Casaburi R, Whipp B (1999) Principles of exercise testing and interpretation: including pathophysiology and clinical applications, 3rd edn. Lippincott Williams & Wilkins, Philadelphia, Appendix C, pp 531–540